Answer to Reviewer 1

We thank the reviewer for thorough reading and thoughtful comments and suggestions. A detailed discussion of the changes that we made in response to the reviewer's comments is given below. In what follows, we state the reviewer's comment in boldface, and describe our response in plain text. Text in the manuscript is represented in italics. The text that has been modified/included in the new version has been highlighted in red.

"Overall this is an excellent manuscript, presenting a new result about the Earth's ocean-land-atmosphere mass exchange, using a unique combination of satellite and reanalysis datasets, and a clear easy-to-follow methodology."

We appreciate the positive overall comment about the manuscript.

"The only major concern/question I have is this: the interbasin ocean transport N is a small residual of differencing large numbers. I see that each set of numbers is followed by a 95% confidence range, and I read without quite understanding that the confidence interval is computed by a bootstrap method on the data itself. I don't believe the re-analysis data have their own error estimates; I believe the GRACE data do but those did not seem to be used in the confidence interval estimation. I wonder whether estimating uncertainties in the transports by propagating uncertainties in the inputs would give intervals consistent with those of the bootstrap method. Upper bounds on the uncertainties in the inputs can be estimated, for example, by comparing UT-CSR mascons to JPL or GSFC mascons, by comparing ECMWF reanalysis to NCEP or another model's reanalyses, etc. I say this because the lack of correlation between the inter- annual transports and ANY index of ocean-atmosphere interaction (ENSO, SOI, etc) is suspicious."

The following changes have been included to address the issues raised by the referee:

1. **Bootstrap:** We have included an intuitive description of the bootstrap method for time series and a reference to a paper on bootstrap method for time series. Besides, we have provided extended details about how confidence intervals have been evaluated:

   The reported 95% confidence intervals and the correlation coefficients are evaluated using the stationary bootstrap scheme of Politis and Romano (1994) (with optimal block length selected according to Patton et al., 2009), and the percentile method. The intuition underlying the bootstrap is simple. Suppose that the observed time series x_1, ..., x_n is a realization of the random vector (X_1,..., X_n) with joint distribution P_n and which is assumed to be part of a stationary stochastic process. Given $\boldsymbol{X}_n$, we first build and estimate $\hat{P}_n$ of P_n. Then B random vectors $(X_1^*, \dots, X_n^*)$ are generated from $\hat{P}_n$. If $\hat{P}_n$ is a good approximation of P_n, then the relation between $(X_1^*, \dots, X_n^*)$ and $\hat{P}_n$ should well reproduce the relation between (X_1,..., X_n) and P_n (for an introduction of bootstrap methods for time series see Kreiss and Lahiri (2012) and the references therein). Here, the number of bootstrap replications was set to B=2000. In general, half length of the confidence interval can be very well approximated by twice the standard deviation of the sample mean estimated from the bootstrap replications. Prior to applying the bootstrap to a time series, least-squares estimated linear/quadratic trend and sinusoid with the most relevant frequencies are removed from it to meet the stationarity conditions of the method. In particular, each series

has been decomposed into trend, seasonal and residual components. The bootstrap is applied to the residual component producing bootstrap samples of the residuals. For the evaluation of confidence intervals for the different components of WT, the trend and seasonal terms are added back (to the bootstrap sample of the residuals) producing bootstrapped time series of the component of interest. These samples are then used for further analysis. As an illustration, for the WT N component we proceed as follows: (i) a model with linear, annual, and semiannual signals is fitted to the data. The fitted linear trend and annual and semiannual signals are subtracted from the original time series; (ii) the stationary bootstrap is then applied to the residuals producing 2000 bootstrap samples of the residuals; (iii) The estimated trend and seasonal components are added back to each bootstrap sample of the residuals obtaining an ensemble of 2000 bootstrapped time series for the N component; (iv) these 2000 bootstrapped time series are used to obtain 95% confidence intervals for the mean fluxes (average of N over the 14 year period of study) and for the amplitude and phase of the annual component using the percentile method. For the mean fluxes, the average of N for each of the 2000 bootstrapped time series was first evaluated and then the 0.025 and 0.975 percentiles of these 2000 averages were reported as 95% confidence interval. For the study of the climatology, a linear trend model with annual and semiannual components was fitted to the 2000 bootstrapped time series producing corresponding estimates of the annual amplitude and phase. The 0.025 and 0.975 percentiles of these estimates were reported as 95% confidence intervals. In order to study the robustness of the results with respect to the model choice, the analysis is rerun using 11 alternative models obtained considering different forms for the trend component (quadratic or constant) and including higher frequencies in the harmonic regression (up to 5). The results are robust. The relative difference with respect to the reported values is smaller than 1.2% for point estimates and smaller than 3.3% for the extremes of the 95% confidence intervals.

2. **Confidence intervals of the correlation coefficients**. More details are provided:

   Note that for the study of correlation the bootstrap was applied to the bivariate time series of the residuals of the two variables of interest producing an ensemble of 2000 bivariate time series of residuals. For each bivariate time series of residuals the correlation between the two components of the series was first evaluated. The average and the 0.025 and 0.975 percentiles of these 2000 estimates were reported as point estimate and confidence limits for the correlation between the two variables of interest (correlation between residual components is used to avoid spurious correlation).

3. **Bootstrap Vs Error propagation:** The confidence intervals estimated from bootstrap have been compared to those estimated from error propagation of the mascon. As CSR mascon solution does not provide such error estimates, we have used the JPL mascon solution for the comparison. An explanation of why bootstrap confidence intervals contains, as expected, the error propagation confidence interval has been also provided. In the description of the bootstrap method we have included the following text:

   As an independent check of the bootstrap, confidence intervals for the mean value of N have been also evaluated by propagating the error estimate in GRACE data (using the JPL GRACE mascon solution for which error estimates are available). The resulting intervals were

consistent with those of the bootstrap method. In particular (see Section 4 for details), we show that in all cases the bootstrap intervals contain the intervals obtained from error propagation. In this respect, the CI_{95} from bootstrap analysis can be considered a conservative estimate. This should be expected, since the residual component underlying the bootstrap approach includes measurement errors and other type of errors (related, for example, with the estimate of the trend and seasonal terms). As a result, the uncertainties in the transports estimated by the bootstrap should be larger than the corresponding uncertainties estimated by error propagation.

We have included a new section 4, entitled "Comparison with other datasets", which includes the comparison between error propagation and bootstrap confidence intervals for the *N* component estimated from JPL data:

CSR GRACE mascon solution is replaced by the JPL GRACE mascon solution provided by the Jet Propulsion Laboratory/NASA (Watkins et al., 2015; Wiese et al., 2019). Similarly to CSR data, JPL are corrected for GIA effects, C_{20} Stoke coefficients are replaced by a solution from SLR, and data are reduced to 1º regular grids from 0.5º regular grids. Besides, we have applied the degree-0 Stoke coefficients correction. However, CSR and JPL mascon solutions are not directly comparable. The main reason is that an estimate of degree-1 coefficients has been added to JPL mascon solutions, and the GAD product has not been added back. The corrections applied by JPL are not supplied separately and we cannot do/undo any of the corrections to process JPL data as we did with CSR data. In particular, the GAD product is not available for JPL. In any case, the JPL solution is useful here since it provides an error estimate of the mascon solution that can be propagated to obtain confidence intervals of N, which are independent from those estimated with the bootstrap analysis. Table 2 shows the CI_{95} of the mean values of the N component for different ocean basin estimated from error propagation and bootstrap analysis. It is observed that in all cases the CI_{95} from error propagation are included in those from bootstrap analysis, meaning that the latter are a conservative estimate of the error. JPL propagated error can be expected to be similar to that propagated from CSR error estimates (which are not available), and then we can assume that the reported CI_{95} for N calculated from CSR data are a conservative estimate. Besides, comparing Tables 1 and 2, it is observed that the mean values of N are quite similar and that the CI_{95} largely overlap. Regarding to the time variability, the values of the N component from CSR and JPL mascon solutions show Pearson correlation coefficients greater than 0.85 (p-value < 10^{-3}), except for the Atlantic (0.70). Thus, despite the different processing of CSR and JPL data, the reported analysis for the N component is robust with respect to the choice of GRACE datasets.

Table 2. Mean net WT from JPL mascon for different ocean basins according to Equation 2 . CI_{95} are estimated as propagation of mascon errors provided by JPL, and from bootstrap analysis. Units are Gt/month.

		Mean (CI_{95} from error propagation)	Mean (CI_{95} from bootstrap)
Outflows	Pacific	1182 (1143, 1220)	1182 (1062, 1306)
	Arctic	735 (713, 757)	735 (711,761)
	Pacific + Arctic	1917 (1872, 1961)	1917 (1806, 2036)
Inflows	AIA	1183 (1092, 1274)	1183 (1077, 1282)
	Atlantic	919 (866, 972)	919 (845, 985)
	Indian	999 (980, 1018)	999 (928, 1067)
	Atlantic + Indian	1918 (1862, 1974)	1918 (1838, 2003)

4. **Other P and E datasets:** According to ERA5 documentation, there exists error estimates. Unfortunately, they are not available for the general public as us. In any case, we have included new computations with several *P* and *E* datasets. It is included in the second point of the new section "Comparison with other datasets":

ERA5 P and E data are replaced by several datasets for comparison purposes. The objective is not to be exhaustive in the selection, but rather to show that the reported features of the N component are quite robust with respect to the choice of the P and E datasets. The data sets considered are:

(i) Continental P from GPCC (Schneider et al., 2011), GPCP (Adler et al., 2018), CMAP (Xie and Arkin, 1997), UDel (Willmott and Matsuura, 2001), and GLDAS/Noah (Rodell et al., 2004; Beaudoing and Rodell, 2016).

(ii) Ocean P from GPCP and CMAP.

(iii) Continental E from GLEAM (Miralles et al., 2011; Martens et al., 2017) and GLDAS/Noah.

(iv) Ocean E from OAFlux (Yu et al., 2008) and HOAPS/CM SAF (Schulz et al., 2009).

The Pacific outflow is estimated with the 162 possible combinations of P and E, including ERA5. The time period is 2003-2016, except for HOAPS/CM SAF and GPCP, which span from 2003 to 12/2014 and 10/2015, respectively. The degree-0 corrections in GRACE data is made for each combination. Note that only ERA5 includes P and E for both continents and oceans. All grids have been homogenized to 1° regular grids. The main concern here is the heterogeneity of the spatial coverage among datasets. To make the results comparable among datasets, the computations are restricted to the common grid points, which do not cover the entire Earth (Figure 8a). However, in spite of the fact that due to the partial coverage the principle of water mass conservation is not accomplished, the Pacific outflow obtained in the common grid points from ERA5 (black line in Figure 8b) is quite in agreement with the same signal obtained with global coverage (red line in Figure 3 which is also reported as red line in Figure 8b). The Pearson correlation coefficient between the two signals is 0.994 (p-values $< 10^{-3}$) with an average difference around 50 Gt/month. In general, the Pacific outflows estimated from all the P and E dataset combinations show qualitatively the same signal than the one reported in Figure 3. For each of the 162 estimates of the Pacific outflows corresponding to the possible P and E dataset combinations, we evaluated the average outflow (over the period of study), which is 968 Gt/month (STD: 489), and the correlation with the Pacific outflows in Figure 3, which is 0.82 (STD: 0.06; p-values $< 10^{-3}$).

These experiments show that the reported net WT are physically consistent among datasets, at least qualitatively."

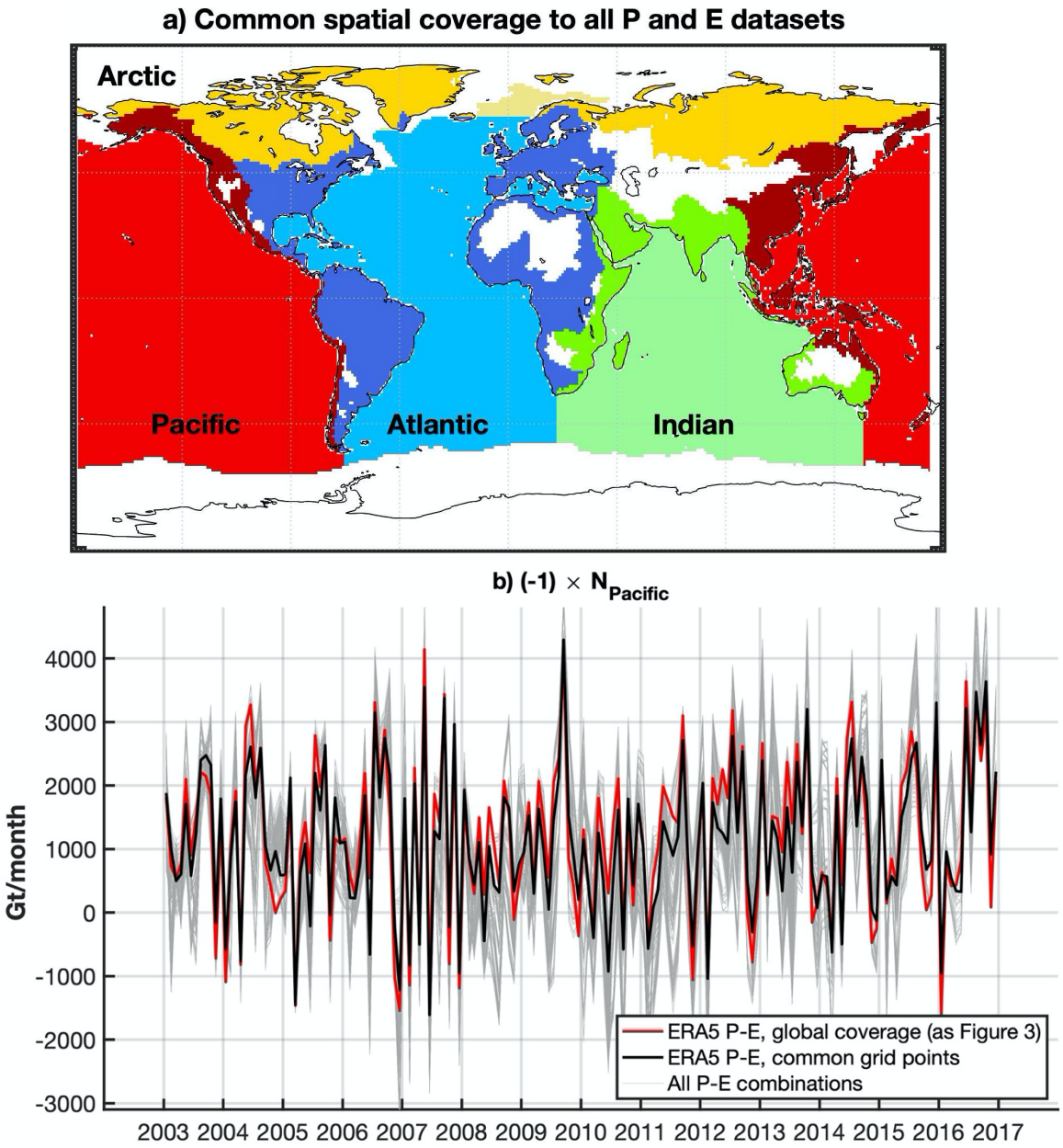

Figure 8. Monthly time series of (the opposite of) the Pacific outflow estimated from 162 combinations of *P* and *E* datasets. a) Spatial coverage common to all datasets. b) Pacific outflows: Gray thin curves are the 162 Pacific outflows estimated in the common grid points to all datasets (no global coverage); black and red curves are based on ERA5 *P* and *E* and are obtained using either only the grid points common to all datasets (black curve) or global coverage (red curve). Note that the red curve is the same as in Figure 3.

5. **Lack of correlation.** We have included a discussion on the lack of correlation between the inter-annual transports and the indices of ocean-atmosphere interaction. In particular we propose the two following explanations:

   "To explore this lack of correlation, we have estimated the correlation coefficient between each climatic index and each WT component (Figure 7b).

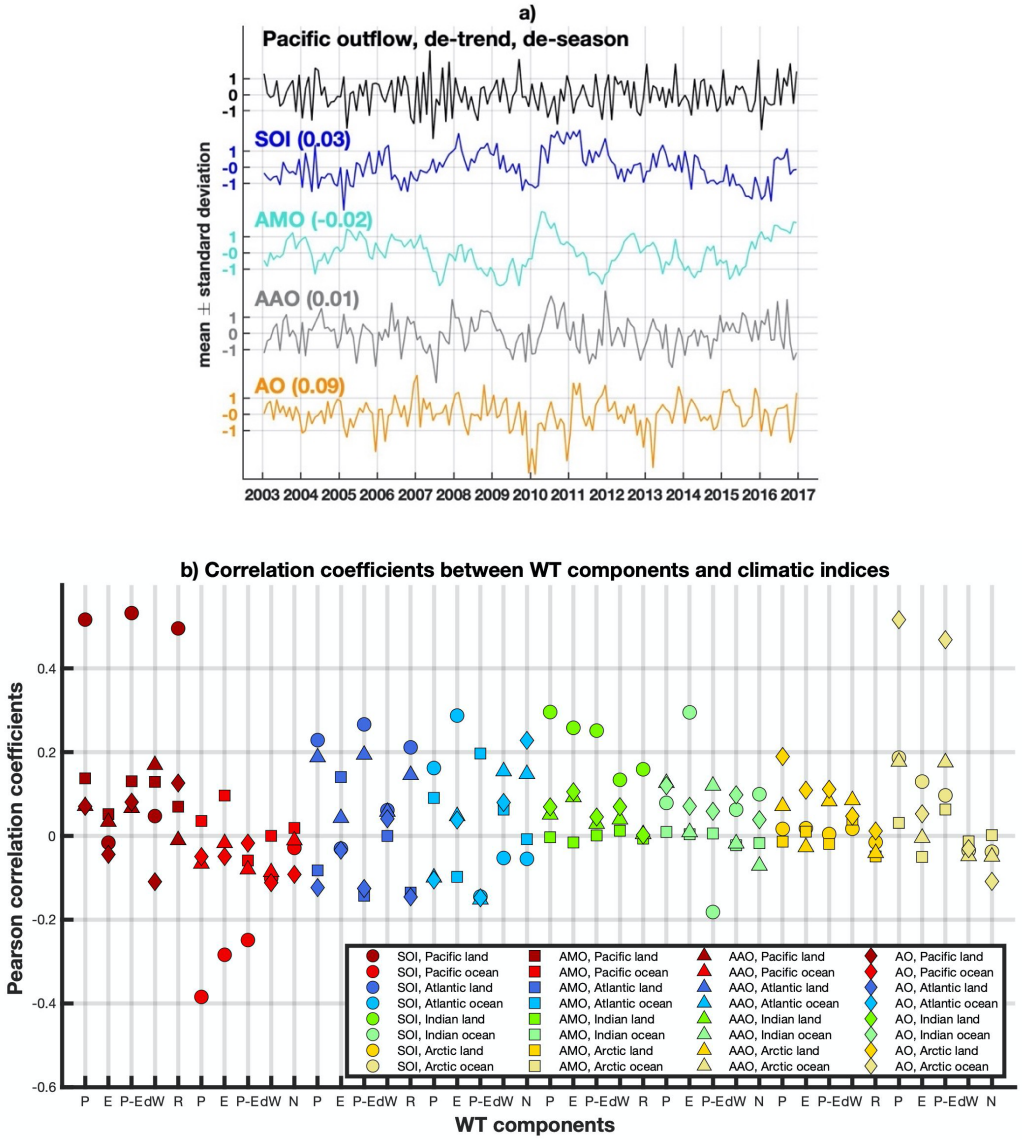

Figure 7. Pacific outflow and climatic indices for ENSO, AMO, AO, and AAO. a) Time series of Pacific outflow is de-trend and de-season. All time series are normalized to have unit variance. Values in the parenthesis are the correlation coefficient between the corresponding climatic index and the Pacific outflow. b) Correlation coefficients between de-trend and de-season WT components of different regions and the climatic indices.

All of them are lower than 0.3 except for 6 cases in 2 regions. In the Arctic, P and P−E in the drainage basins of the Arctic show a correlation of ~0.5 with the AO. This correlation is natural since that is the area of influence of the AO. The other region is the Pacific, where, as expected, the SOI shows a correlation around 0.5 with P, P−E, and R in the drainage basins, and around −0.4 with P in the ocean. However, this individual correlation does not extend to the Pacific outflow. In order to understand why this is the case, it is convenient to express the N component of the water transport as a function of (P-E) and dW. According to Equations 1 and 2 we have:

$$N = -(P-E)_{ocean} - R + dW_{ocean} = \underbrace{-(P-E)_{ocean}}_{X_1} \underbrace{-(P-E)_{land}}_{X_2} + \underbrace{dW_{land}}_{X_3} + \underbrace{dW_{ocean}}_{X_4} \qquad (3)$$

It can be shown that the correlation between N and a given index can be express as follows

$$corr(N, Index) = \sum_{i=1}^{4} corr(X_i, Index) \cdot \frac{std(X_i)}{std(N)}, \qquad (4)$$

where corr denotes the correlation coefficient, and std stands for standard deviation. As shown in equation (4), the correlation between N and a given index is a linear combination of the correlation between each component and the index. The coefficients of the linear combination std(X_i)/std(N) are proportional to the standard deviation of each component. The components of equation (4) for the Pacific outflow and the SOI index are shown in Table 3. Despite the fact that some of the individual component exhibits significant correlation with SOI (in particular P−E in land and ocean) when combined with the corresponding coefficients their effects are canceled out yielding to a negligible correlation between water transport and SOI (below 0.03 in magnitude).

Another possible reason for the lack of correlation resides in the definition of the studied regions, for which the presence of subregions with positive and negative influence of an index results in an overall negligible/attenuated influence of the index in the overall region. For example, a positive phase of the AMO is related to an increase of P in western Europe (Sutton and Hodson, 2005), and the Sahel (Folland et al., 1986; Knight et al., 2006; Zhang and Delworth, 2006; Ting et al., 2009), but to a decrease of P in the U.S. (Enfield et al., 2001; Sutton and Hodson, 2005), and northeast Brazil (Knight et al., 2006; Zhang and Delworth, 2006). All these regions are included in the Atlantic drainage basin, and then the influence of a positive phase of the AMO is attenuated."

Table 3. Correlation coefficients between SOI and de-trend and de-season WT components involved to estimate the Pacific outflow according to Equations 3 and 4.

	$std(X_i)$ (Stand. Deviation)	$corr(X_i, SOI)$ (Correlation between X_i with SOI)	$\frac{std(X_i)}{std(N)}$ (Coefficients)	$corr(X_i, SOI) \cdot \frac{std(X_i)}{std(N)}$ (Correlation ·Coefficient)
$X_1 = -(P-E)_{ocean}$	605	0.25	0.57	0.14
$X_2 = -(P-E)_{land}$	212	-0.53	0.20	-0.11
$X_3 = dW_{land}$	96	0.048	0.09	0.004
$X_4 = dW_{ocean}$	711	-0.10	0.67	-0.07
Corr(*N*,SOI)				-0.03

Note that table 3 provides also some insights about the causes of the interannual variability of Pacific Ocean outflow. The largest standard deviation of *P−E* and *dW* in the ocean suggests that these two components might drive the interannual variability of the Pacific Ocean outflow. This is confirmed by a correlation analysis. The correlation between *N* and the $(P-E)_{Ocean}$ is -0.70. The correlation between *N* and the dW_{ocean} is 0.84. The correlation of *N* with the corresponding land components is below 0.18. In all cases, prior to the evaluation of the correlation the corresponding time series have been de-trend and de-season.

Now addressing some details:

(1) Figure 1: I would have liked to see a row with P-E- R next to the row for dW in Figure 1.

Figure 1 probably means Figure 2. Including *P-E-R*, in our opinion, is not very useful since, by definition of *R*, *P-E-R* will perfectly match *dW*. The comparison would be interesting with an independent dataset of R.

(2) Figures 1 and 3: I am sure the authors know better smoothers than the running mean (Hanning, Kaiser, etc). I recommend they use one.

We have replaced the running mean by a low pass filter defined by a Hann function of 24 months (the resulting smoothed curve is quite in agreement with the one previously obtained by running mean smoothing)

(3) Line 27: Clark reference missing. Recheck all your references, I did not do an exhaustive check.

Thank you. We have checked all the references.

(4) Line 93: tectonic signals in the gravity field do not 'masquerader as mascons'. Mascons are a simple mathematical representation of the gravity field with a physical interpretation. Tectonics "would be incorrectly interpreted as water mass flux"

Thank you. It is better expressed in this way. We have re-written the sentence: *"Any other non-surficial effect such as long-term tectonics would be incorrectly interpreted as water mass fluxes…"*

(5) Lines 124 et seq: see my concern above. A physical interpretation of this mathematical approach to confidence intervals would be useful.

We have extended the description of the bootstrap - see point 3 (Bootstrap Vs Error propagation) in page 1 of this response.

(6) Line 164: and loses 'to the atmosphere' 879 Gt/month. . .

The sentence has been re-written:
"On average, the Atlantic Ocean receives 926 Gt/month (CI_{95}=[876, 980]; or 0.36 Sv) of salty water, and loses to the atmosphere 879 Gt/month (CI_{95}=[828, 930]) via P−E+R."

(7) Line 188: I think 'The Atlantic/Arctic inflow 'mirrors this behaviour' is a better phrase in English.

Thank you. We have re-written the sentence: *"The Atlantic/Arctic inflow mirrors this behaviour."*

(8) Somewhere: W. T. Liu et al (GRL 2006, on South American water balance) did a similar estimation of water flux between an ocean basin and the land, without using any numerical model data.

Thank you. We agree that it is a pertinent reference. We have included it in the last paragraph of the introduction, which now is:

"In this work we propose a new methodology devised to estimate the net WT through the boundaries of a given oceanic region. A defining feature of the proposed approach is the use of the time-variable gravity data from the GRACE (Gravity Recovery and Climate Experiment) satellite mission to estimate the change of water content. We apply the methodology, in conjunction with conventional meteorological data of general hydrologic budget schemes, to estimate the time evolution over the period 2003-2016 of the net WT and exchanges among the four major ocean basins – namely Pacific, Atlantic, Indian, and Arctic. We analyse and report our results of the seasonal climatology as well as the interannual variability of WT. Such information, not available previously, is of valuable importance. For example, in closed regions, net WT through the boundaries on the surface must be counteracted by moisture fluxes through the same boundaries in the atmosphere to match GRACE measurements. Such approach has been successfully applied to study the hydrological cycle of South America (Liu et al., 2006). At ocean basin scale, knowledge about net WT not only would help elucidate the role of the oceans within the water cycle, but it will also impose restrictions on moisture advection in the atmosphere that would help to improve atmospheric models. On the other hand, ocean models usually deal with inflows and outflows of a given ocean region (Warren, 1983; Rahmstorf, 1996; Emile-Geay et al., 2003; de Vries and Weber, 2005; Dijkstra, 2007). Net WT estimates for such ocean region would be useful to impose constraints to the relationship between its inflows and outflows, which would improve the reliability of the models. Better models will improve our knowledge of the Earth's WT dynamics and its evolution in the future, which is critical in the present scenario of climate change."

(9) There are a few more minor language errors (lines 255, 267 and possibly others). Please go over the manuscript and clean up.

Done. Thank you.